MVA2002 IAPR Workshop on Machine Vision Applications, Dec. 11 - 13, 2002, Nara- ken New Public Hall, Nara, Japan

8—19

EXTRACTION OF NON MANUAL FEATURES FOR VIDEOBASED SIGN LANGUAGE RECOGNITION

Ulrich Canzler[1]
Chair for Technical Computer Science
Aachen University Of Technology

Thomas Dziurzyk[2]
Chair for Technical Computer Science
Aachen University Of Technology

Abstract

Videobased sign language recognition is barely investigated in the field of image processing. General conditions like realtime-ability and user- and environment– independence require compromise solutions. This paper presents a system for automatic analyzing of the facial actions. For this point distribution models and active shape models are brought into action. Additional a comparison is made between different approaches for the shapes initialization.

1 Introduction

Sign language is the natural language of deaf people. It is a non-verbal and visual language, different in form from spoken language, but serving the same function.

It is characterized by manual parameters (hand shape, hand orientation, location, motion) and non-manual parameters (gaze, facial expression, mouth movements, position, motion of the trunk and head).

Especially the non-manual parameters can be crucial, for example the mouth movements are coding several functions: They specify meanings of a sign (meat / hamburger), emphasize details, they make ambiguous signs using same manual parameters well defined (brother/sister) and supply the recognition of a sign at all by giving redundant information.

Existing sign language recognition systems rely exclusively on the manual parameters [1] [2] [3] although image processing offers the possibility to consider non-manual parameters, too.

This work presents a coarse overview of a system in development for the automated analysis of the human mimic. The extraction of eye- and lip features using a biomechanical model is described in detail.

2 Methodology

The system consists of a ***face finder and –tracker*** module, that combines several probability maps, analyzing motion by temporal templates and skin color [4] by a RGB-histogram, and classifies many geometric features [5] (relation height/width, orientation, roundness, invariant moments etc) and abstract features described by Jones [6] with an Adaboost classifier. The overlaying of the four maps results in a very accurate bounding box margin the face. The scene clipping of the cam becomes optimized for the following processing levels by tilting and panning the cam and finally shifting the zoom-lens.

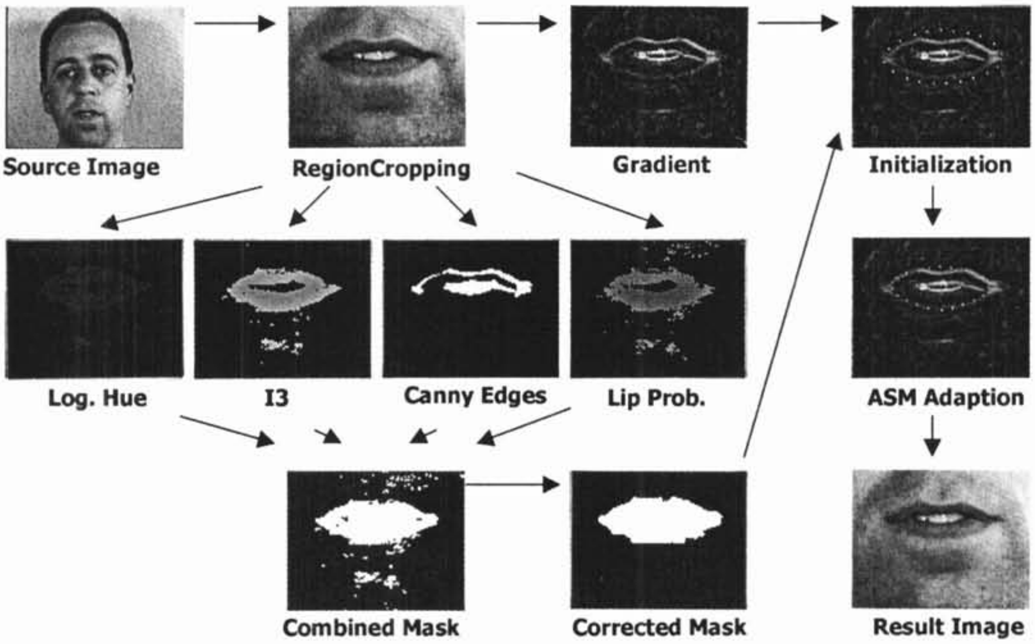

fig 1: System overview

[1] Address: 52062-Aachen, Germany. E-mail: **canzler@techinfo.rwth-aachen.de**

[2] Address: 52062-Aachen, Germany. E-mail: **dziurzyk@techinfo.rwth-aachen.de**

The next task to do is the finding and extraction of non-manual features. Each of this features must be investigated on a different way. E.g. the ***analysis of the gaze*** needs simple template matching techniques on the gradient of the eye region. For this first a eye-color histogram was created by 520 manual segmented images. By using the integral projection and this adaptive eye-color histogram we receive very accurate results for eye-position and -tracking.

Applying this method for ***finding the nostrils*** we get further characteristic points, that we map on a biomechanical model. This 3d-model of the human face allows to predict the muscle's tensions, by simulating the skin and the muscles with springs, so it serves for additional verification. E.g. a smile tends to return into a base-lined expression.

Lip movements are more complicated to parse. Lip corners are permanent visible features in frontal views, so it is easy to find them by performing a horizontal and vertical integral projection inside a cropped area. Additional we use the result of four feature maps, described below. This information combined with a Susan edge detector yields excellent initial positions for a Point Distribution Models (PDM), that represents the shape and its possible deformation of the lip outline.

Active Shape Models (ASM) align this PDM shapes and correct invalid shape matching. So they are good suited for representation of an object like human lips as a set of N labeled landmark points in a vector. The ASM module is described later in this paper, first we address the initialization problem.

To get rid of lightening conditions, we correct the appearance of the mouth region by shifting the HSI-channels to an optimal mean value. The color channel can be used additionally to neglect color discrepancies between the images. In our database you can notice, that there are significant differences reg. the lip- and skincolor. caused by illumination and person dependency. This affects negatively the creation of the maps, which are based on the red and yellow channel in the RGB color space. For this reason the mouth region in each image get calibrated to before trained mean values. All occurrences of the colors are counted and the average value on the limited range is determined.

Next for all pictures a common average value was selected, which should represent the natural skin color.

All occurrences of the colors are counted and the average value on the limited range is determined. Next for all pictures a common average value was selected, which was to show the natural skin color. Now the average value can be computed for each picture on its cutout. If this deviates from the selected average value, for all pixels of the picture the color values are shifted accordingly.

In order to initialize the Active Shape Model as good as possible, for the initialization mask we tried to approximate the lip outlines and the lip corners on the basis of the different characteristics. For this we use four maps. Three of these maps use colored separation characteristics of lips and face.

The first map uses the Bayes theorem over a histogram, which was provided for lip-similar colors. Thus the pixel probabilities of affiliation can be determined. By a threshold value the pixels are divided in two groups, i.e. in lips and face.

$$p(lip/rgb)=\frac{p(rgb|lip)p(lip)}{p(rgb|lip)p(lip)+p(rgb|\neg lip)p(\neg lip)}$$

The second map is basing on the special condition of the lip color. This contains many red and green colours into first line. The combination of the two colors and the thresholding following on it supplies a map, which emphasizes pixels of the lip colors.

The third map is won from the combination of the 3 RGB channels, whereby a special weighting is selected here. The HSV area is characterized by complex transformations on the one hand and singularities and saltuses on the other hand. Therefore we search for a color space, which fulfills the following criteria:

- No saltuses or singularities, which make a segmenting more difficult of the data
- Good separation from brightness and color information.
- Simplification of segmenting by a separating barness of matching color regions as good as possible.
- Simple and fast conversion of and/or in RGB space.

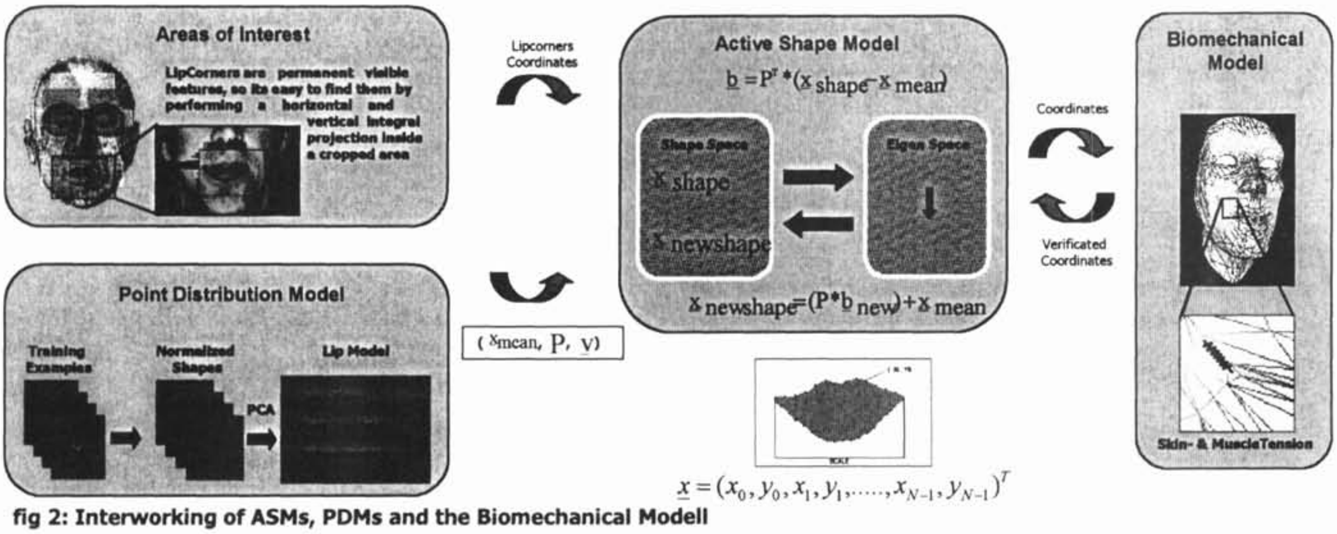
Areas of Interest
LipCorners are permanent visible features, so its easy to find them by performing a horizontal and vertical integral projection inside a cropped area
Lipcorners Coordinates
Point Distribution Model
Training Examples
Normalized Shapes
PCA
Lip Model
(x_{mean}, P, $\underline{v}$)
Active Shape Model
$\underline{b} = P^T * (\underline{x}_{shape} - \underline{x}_{mean})$
Shape Space
$\underline{x}$ shape
$\underline{x}$ newshape
Eigen Space
$\underline{x}_{newshape} = (P * \underline{b}_{new}) + \underline{x}_{mean}$
$\underline{x} = (x_0, y_0, x_1, y_1, \ldots, x_{N-1}, y_{N-1})^T$
Coordinates
Verificated Coordinates
Biomechanical Model
Skin- & MuscleTension

fig 2: Interworking of ASMs, PDMs and the Biomechanical Modell

For the lip segmenting particularly the modified I3-channel was found to be useful. Since the blue channel for the lip color plays a subordinated role, it turned out that its restriction leads to better segmentation results. The following transformation for the I 3 channel was selected:

$$I_3 = \frac{2G - R - \frac{B}{2}}{4}$$

For the ***fourth map*** a gradient picture won by the Canny operator was used, whereby a Gaussian filter smoothed the edges.

The four characteristic maps are combined by a Bayesan Belief Network to a result map and corrected afterwards to one initialization mask; disturbances are settled and closed with the morphological operators dilatation and erosion.

Finally the algorithm supplies an object, which matches the contour in good approximation at the material outlines of the lips and serves for the ASM-initialization as well as verifying the lip corner positions retrieved by the integral projection described above.

The last module of the system deals with the problem of the feature analysis, which has the objective of classifying the described features to so-called Action Units (AU's). This AU's are defined as basic anatomically deduced minimal movement units of the face.

We use a System derived from the Facial Action Coding System (FACS) [9] to describe this movements. Each AU can be performed with three different intensity levels, simultaneous occurrences of several AU's are permitted. The AU's themselves are won from the above described features by using a classifier, based on Hidden Markov Models.

To improve the classification, additional rules are applied by Fuzzy Sets for effect concerning dominance, substitution and exchangeability. This way additional contextual knowledge is taken into account. Finally the system yields the facial expressions coded by the AUs. This high level features are afterwards combined with the manual-parameters extracted from a separate system developed at our department.

3 Results

For testing the system, we create a database with 720 images (24 persons, 30 images under different lightning conditions and mouth openings). We defined two qualities for the recognition rates. The first one defined a maximum distance between hand segmented and automatic explored points on the lip contour. The second one was more tolerant with 6 pixels displacement.

As result it can be held that the ranges within 3 pixels were usually correctly detected around the upper and/or lower lip, which confirmed also the visual observation. Here the detection rate was about 80%, whereas the recognition rate around the left and right lip corner with the asymmetrical PDM was ca. 60% and by using the symmetrical PDM by ca. 53% correctly recognized.

Total regarded it can be stated that in most cases the errors took place within the lips and only rarely outside, then however mostly with persons with beard, with whom the strong gradient drew the points in wrong direction.

Thus the upper and/or lower lip outline was detected with approx. 15% of the pictures below and/or above the actual upper/lower lip outline. The lip outline within the range of the right and left lip corners was determined in each case with approximately 30% of the images on the left of and/or right by the actual lip outline. A rise of the error barrier on six pixels led both for the symmetrical and asymmetrical PDM in each of the four lip ranges to a correct detection of at least 94%.

Exact results show the tables 1; it differs between the used PDMs (symmetrical or asymmetrical) and the maximal tolerated distance (3 or 6 pixels) with respect to the appropriate mouth ranges and shows the direction of the shifted points.

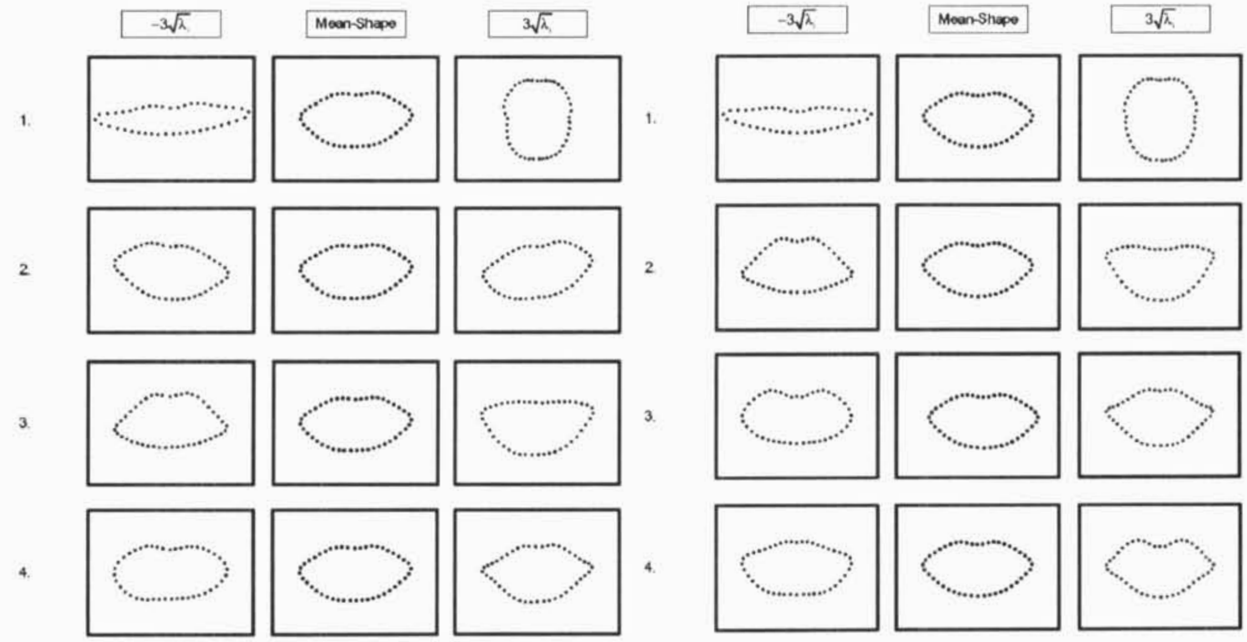

fig 3: Deformations of the lip PDM (left: asymmetrical shapes, right: symmetrical shapes

Deviation	≤ 3 Pixel				≤ 6 Pixel			
Area	Up	Right	Down	Left	Up	Right	Down	Left
Number								
←	5	256	22	5	0	28	0	0
→	46	4	9	226	3	0	1	31
↑	15	7	76	90	0	0	1	3
↓	105	138	10	4	13	9	1	0
correct	549	315	603	395	704	683	717	686
in %								
←	0,69	35,28	2,92	0.69	0,00	3,89	0,00	0,00
→	6,39	0,56	1,25	31,25	0,42	0,00	0,14	4,31
↑	2,08	0,97	10,56	11,94	0,00	0,00	0,14	0,42
↓	14,58	18,47	1,39	0,56	1,81	1,25	0,14	0,00
correct	76,25	43,75	83,75	54,86	97,78	94,86	99,58	95,28

Deviation	≤ 3 Pixel				≤ 6 Pixel			
Area	Up	Right	Down	Left	Up	Right	Down	Left
Number								
←	10	253	16	6	0	31	0	1
→	14	7	15	204	2	2	2	27
↑	16	20	72	14	0	0	0	0
↓	112	22	7	56	14	2	1	8
correct	568	418	610	440	704	685	717	684
in %								
←	1,39	34,72	2,22	0,83	0,00	4,31	0,00	0,14
→	1,94	0,97	2.08	28,19	0,28	0,28	0,28	3,75
↑	2,22	2,78	10,00	1,94	0,00	0,00	0,00	0.00
↓	15,56	2,92	0,97	7,78	1,94	0,28	0,14	1.11
correct	78,89	58,06	84.72	61,11	97,78	95,14	99,58	95,00

tab 1: Recognition rates (left: asymmetrical shapes, right: symmetrical shapes

4 Conclusion

In this paper we presented a system for extracting facial expression features for sign language recognition. Several modules are necessary for detecting, tracking and analyzing the face and finally to create a feature vector. We decided to work out an approach with Active Shape Models for modeling the lip movement. Symmetrical and Asymmetrical Shapes delivered nearly same recognition rates.

5 Acknowledgement

This research was supervised by Prof. Karl-Friedrich Kraiss, Department for Technical Computer Science at the RWTH-Aachen and supported by the DFG-Project SIGMA.

References

[1] H. Hienz, K.-F. Kraiss, B. Bauer. Continuous Sign Lan guage Recognition using Hidden Markov Models. Pro ceedings of the Second International Conference on Mul timodal Interfaces, Hong Kong (China), 1999.

[2] T. Starner and A. Pentland. Real-time American Sign Language recognition from video using hidden Markov models. Perceptual Computing Section Technical Report o. 375, MIT Media Lab, Cambridge, MA, 1996.

[3] C. Vogler and D. Metaxas. ASL recognition based on a coupling between HMMs and 3D motion analysis. CIS Technical Report, Department of Computer and Informa tion Science, University of Pennsylvania, 1997

[4] M. Jones, J. Rehg: Statistical color models with appli cation to skin detection, Proceedings Computer Vision and Pattern Recognition, 1999, pp. 274-280

[5] S. Gong, S. McKenna, A. Psarrou: Dynamic Vision, Imerial College Press, London, 2000

[6] P. Viola , M. Jones: Robust Real-time Object Detection, Technical Report Series, Cambridge Research Laboratory, Feb 2001

[7] P. De Smet and R. Pries. Implementation and analysis of an optimized rainfalling watershed algorithm. In IS&TSPIE's 12th Annual Symposium Electronic Imaging 2000: Science and Technology Conference: Image and Video Communications and Processing, San Jos, Califor nia, USA, January 2000. 759–766.

[8] D. Comaniciu and P. Meer. Robust analysis of feature spaces: Color image segmentation. In Proc. IEEE Conference on Computer Visiona and Pattern Recognition, Puerto Rico, 1997. 750–755.

[9] P. Ekman, W. Friesen: Facial Action Coding System, Consulting Psychologists Press Inc., California, 1978